Analiza sentimentelor din spatele recenziilor curselor cu uberul

Uber Ride Reviews Dataset

# Introducere

Pe masura ce tehnologia avanseaza totul devine tot mai digital si automatizat. Astazi putem cumpara totul de pe site-urile specializate, iar serviciile pot fi comandate online cu un singur click. Dar sa judecam un produs doar dupa imaginile si recenziile cilentilor care au cumparat déjà produsl respective este destul de dificil si necesita o analiza amanuntita deoarece, pentru un produs sai serviciu exista de obicei, sute sau poate mii de recenzii.

Analiza sentimentelor este cea mai proeminenta ramura a procesarii limbajului natural si se ocupa de clasificarea textului pentru a determina intentia autorului. Intentia acestuia poate fi de tip admirational(pozitiva) sau critic(negativa). In aceasta lucrare, imi propun sa analizez recenziile curselor cu uberul stocate in baza de date Uber Ride Reviews Dataset si, folosind metode Data Mining sa implementez un model care sa clasifice o recenzie noua in functie de intentia autorului, drept negativa sau pozitiva.

# Literatura de specialitate

Fiind un domeniu vast, analiza sentimentelor a fost studiata de mai multi cercetatori din intreaga lume care au incercat sa construiasca metode de clasificare a recenziilor produselor sau serviciilor care sa aiba o acuratete cat mai mare.

In lucrarea “*Classification* *of Sentimental Reviews Using Machine Learning Techniques*” scrisa de Abinash Tripathy, Ankit Agrawal si Santanu Kumar Rath este prezentata o comparatie a rezultatelor obtinute prin aplicarea algoritmului de clasificare Naïve Bayes(NB) si Suport Vector Machine (SVM). Setul de date folosit pentru instuirea si testarea modelelor in aceasta lucrare a cuprins recenziile filmelor si s-a facut o comparatie a rezultatelor obtinute cu cele existente déjà in literaturade specialitate. Rezultatul studiului a aratat faptul ca metoda SVM depaseste orice alt clasificator in prezicerea sentimentului unei recenzii. O alta abordare a clasificarii recenziilor este propusa de Deepali Virmani, Preeti Arora si Pradnya Satish Kulkarni in articolul “ *Cross domain analyzer to acquire review proficiency in big data*”. Acestia propun o clasificare transversala a domeniilor, un analizator de domenii incrucisate(CDA) ce extrage cuvintele pozitive din recenziile unui domeniu si le inlocuieste cu sinonimele sale, astfel incat sa reuseasca sa analizeze sentimentele din recenziile unui alt domeniu.Datele utilizate in acest experiment au fost culese de pe Amazon si s-a obtinut o precizie de 92%, algoritmul creand o rafinare de 5% fata de tehnicile deja existente. In literature de specialitate s-a studiat si calitatea produselor software. Deoarece modelele de masurare a calitatii software-ul in uz (QinU) sunt ineficiente si multe dintre ele sunt subiective, Issa Atoum propune in lucrarea intitulata *“ A novel framework for measuring software quality-in-use based on semantic similarity and sentiment analysis of software reviews*” un cadru nou de lucru numit QinUF pentru a masura calitatea software-ului in utilizare folosind recenziile obtinute de aceste produse. Acest nou cadru este alcatuit din trei componente: predictia QinU, clasificarea polaritatii si notarea QinU. In viata de zi cu zi, QinUF permite utilizatorilor sa compare sis a ofere recenzii produselor software. Experimentele effectuate pe parcursul acestui studio au demonstrat faptul ca QinUF a obtinut cea mai inalta performanta, in comparative cu metodele cunoscute. Tot pentru analiza sentimentelor din spatele recenziilor Shahana P.H si Bini Omman folosesc in studiul lor intitulat “ *Evaluation of Features on Sentimental Analysis*” folosesc tehnici precum Mutual Information, Chi-Square, Information gain si TF-idf pentru a selecta caracteristicile din baza de date. Aceste modele sunt evaluate pentru un set de date ce contine 2000 de recenzii ale celor mai cunoscute filme si se investigheaza care este cea mai buna metoda de analiza a sentimentelor din recenzii. Autorii au constatat faptul ca cea mai buna metoda de extragere a sentimentelor din recenzii este unigramul care ofera o precizie de 83% in ceea ce priveste clasificarea sentimentelor.

Deoarece pentru un singur produs sau serviciu pot exista sute sau chiar mii de recenzii si este dificil sa citim si sa analizam toate recenziile pentru a lua o decizie cu privire la achizitionarea acestora,  [Aarti Potdar,](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!)  [Pranav Patil,](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) [Raunak Bagla,](https://www.sciencedirect.com/science/article/pii/S1877050916000983" \l "!) [Rohitashwa Pandey](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) si Prof. [Nagesh Jadhav](https://www.sciencedirect.com/science/article/pii/S1877050916000983#!) ne propun in lucrarea “*SAMIKSHA - Sentiment Based Product Review Analysis System*” un bot de recenzie care genereaza o rezumare faptica a tuturor recenziilor utilizatorilor. Software-ul propus produce o evaluare numerica medie a tuturor caracteristicilor specific produsului pentru a ajuta cumparatorul sa obtina o imagine de ansamblu detaliata a produsului. Astfel, “Samiksha” s-a dovedit a fi un mediu convenabil de analiza a tuturor recenziilor si este benefic pentru cumparatorii de pe site-urile de comert electronic.

Deoarece exista numerice studii pe acesta tema, voi incerca sa aplic o parte din algoritmii si metodele de analiza a sentimentelor din recenzii descrise in aceste articole si voi compara rezultatele obtinute cu cele existente in literature de specialitate.